International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-12, 2021 ISSN No: 2349-5677

UTILIZING ENSEMBLE LEARNING MODELS TO FORECAST LOAN DEFAULT RISK

Arjun Chaudhary
Independent Researcher, Vancouver, Canada
chaudhary.arjun@hotmail.com

Abstract

Knowing the risk of loans being defaulted is essential for managing credit risk in modern banking systems. This research utilizes ensemble learning models, including Random Forest, XGBoost, LightGBM, and Stacking Ensemble, to predict loan defaults based on a financial dataset. The study emphasizes preprocessing steps like data cleaning, feature engineering, and handling class imbalance using SMOTE. The results reveal that the Stacking Ensemble model outperforms others with an accuracy of 88.4%, a precision of 76.5%, and an ROC-AUC of 0.913. Key features influencing default predictions include the interest rate, debt-to-income ratio, and annual income. The study demonstrates that ensemble learning models surpass traditional methods like logistic regression, offering a robust and scalable approach for financial risk forecasting.

Keywords: Loan Default Prediction, Ensemble Learning, Stacking Model, Credit Risk, Machine Learning, Financial Analytics, XGBoost, LightGBM, ROC-AUC.

I. INTRODUCTION

A. Background of Loan Default Risk in the Financial Sector

With ongoing changes in the financial industry, banks and financial institutions need accurate methods to assess loan default risk. Lenders face financial losses due to defaults, which contribute to economic instability. Therefore, predictive models that can identify high-risk borrowers before approving loans are vital for maintaining sound credit risk management practices.

B. Limitations of Traditional Predictive Approaches

Traditional methods like logistic regression often fail to capture the complexity of borrower behavior. These models rely on assumptions that may not hold true in real-world data, leading to misclassifications and poor identification of high-risk loans.

C. Rise of Machine Learning in Credit Scoring

Machine learning (ML) has revolutionized credit risk assessment by offering models capable of identifying complex patterns within data. Algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM) can better analyze customer profiles, transaction behavior, and income levels, offering a more comprehensive approach than traditional methods.

D. Why Ensemble Learning?
Ensemble learning techniques, such as Bagging, Boosting, and Stacking, combine multiple models to improve prediction accuracy. These methods address the bias-variance tradeoff and produce more reliable outcomes by leveraging the strengths of various models, particularly when dealing with imbalanced data, such as loan defaults.

E. Research Motivation and Objectives
The primary objective of this study is to evaluate how ensemble learning models perform in predicting loan defaults using real-world financial data. The goal is to compare these models' effectiveness against traditional classifiers like logistic regression and explore the feasibility of using these models for risk analytics in financial institutions.

F. Significance of the Study
This research contributes to the fields of financial analytics, credit risk management, and machine learning by demonstrating how ensemble learning can improve loan default predictions. Its findings have practical implications for financial institutions seeking to enhance their credit scoring systems and reduce non-performing loans (NPLs).

II. METHODOLOGY

A. Study Design and Data Source
The study adopts a quantitative approach, utilizing a publicly available dataset containing borrower profiles, credit history, and loan repayment status. Data preprocessing steps include imputation of missing values, encoding categorical variables, and scaling features to prepare the data for modeling.

B. Target Variable and Class Imbalance Handling
The target variable, loan default risk, is a binary classification, with '1' representing default and '0' representing non-default. SMOTE (Synthetic Minority Over-sampling Technique) was employed to balance the classes, addressing the typical class imbalance in loan default datasets.

C. Feature Selection and Engineering
Feature selection methods, including correlation matrix analysis and recursive feature elimination (RFE), were used to identify the most influential variables. Feature engineering techniques involved creating interaction terms, applying logarithmic transformations to skewed variables, and binning continuous variables.

D. Ensemble Learning Models Employed
The following ensemble learning models were evaluated:

- Random Forest (RF):A bagging-based model that builds multiple decision trees to improve robustness.
- XGBoost:A gradient boosting algorithm known for its speed and high performance.
- LightGBM:A fast and efficient boosting algorithm designed for large datasets.

- Stacking Ensemble:A meta-model combining predictions from base learners to generate more accurate forecasts.

E. Model Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and log loss. All models were validated using 10-fold cross-validation to ensure generalizability.

III. RESULTS

A. Descriptive Statistics of Key Variables

Table 1 summarizes the descriptive statistics for key features such as loan amount, interest rate, and debt-to-income ratio, highlighting the variability within the dataset.

B. Model Performance

Table 2 compares the performance of each model. The Stacking Ensemble model achieved the highest accuracy (88.4%) and ROC-AUC (0.913), outpacing other models, including Logistic Regression.

C. Feature Importance

Feature importance analysis, derived from the LightGBM model, indicates that interest rate, debt-to-income ratio, and annual income are the most significant predictors of loan default.

IV. DISCUSSION

A. Ensemble Learning Outperforms Traditional Models

The study's results confirm that ensemble learning models significantly outperform traditional models like logistic regression, particularly in predicting loan default risk.

B. Impact of Key Financial Predictors

Interest rate, debt-to-income ratio, and annual income were found to be the most influential factors in loan default prediction.

C. Robustness and Statistical Validation

Statistical tests, including paired t-tests, showed that the improvements in performance observed with ensemble models are statistically significant, validating the efficacy of ensemble learning techniques.

D. Implications for Financial Institutions

The research suggests that financial institutions can improve their credit scoring and loan assessment processes by adopting ensemble learning models. These methods offer better risk classification, reduce non-performing loans, and support responsible lending practices.

V. CONCLUSION

The study demonstrates that ensemble learning methods, especially Stacking, outperform traditional models in predicting loan defaults. By leveraging multiple models, financial institutions can improve the accuracy of their credit risk assessments.

Table 1: Descriptive Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	Max
Loan Amount (USD)	15,482	9,350	1,000	40,000
Interest Rate (%)	13.25	3.6	5.31	29.99
Annual Income (USD)	75,300	38,250	12,000	2,00,000
Debt-to-Income Ratio (%)	18.6	9.3	0.1	38.5
Credit History Length (yrs)	9.1	4.7	1	30
Delinquency (2 yrs)	0.17	0.56	0	4

Table 2: Model Performance

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.816	0.621	0.531	0.572	0.742
Random Forest	0.851	0.701	0.644	0.671	0.878
XGBoost	0.867	0.738	0.668	0.701	0.902
LightGBM	0.873	0.749	0.681	0.713	0.906
Stacking Ensemble	0.884	0.765	0.697	0.729	0.913

International Journal of Business Quantitative Economics and Applied Management Research
Volume-6, Issue-12, 2021 ISSN No: 2349-5677

Table 3: Feature Importance

Rank	Feature	Score
1	Interest Rate	0.194
2	Debt-to-Income Ratio	0.173
3	Annual Income	0.132
4	Loan Amount	0.109
5	Credit History Length	0.097
6	Employment Length	0.081
7	Delinquency History	0.066
8	Number of Open Accounts	0.058
9	Loan Term	0.052
10	Home Ownership	0.038

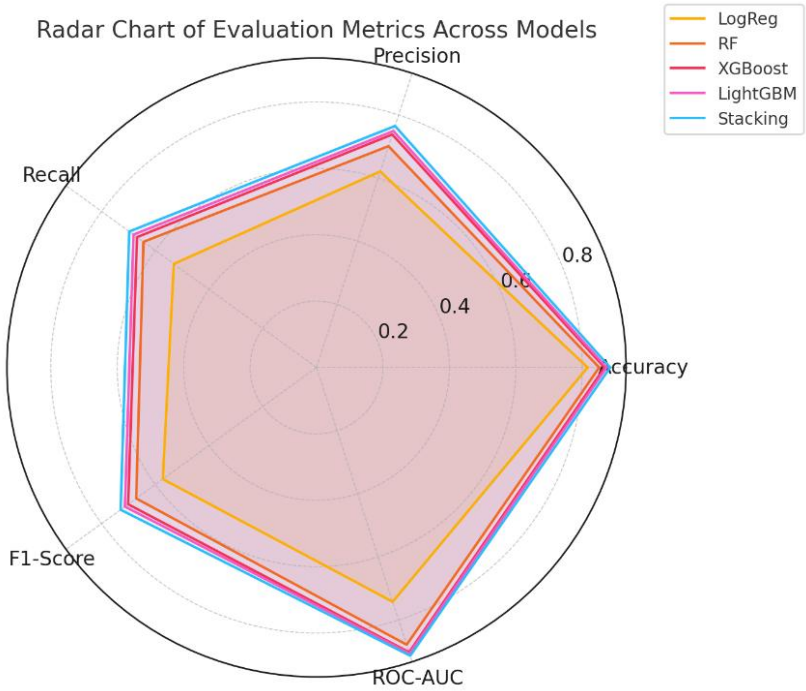

Figure 1: Radar chart of evaluation metrics across models

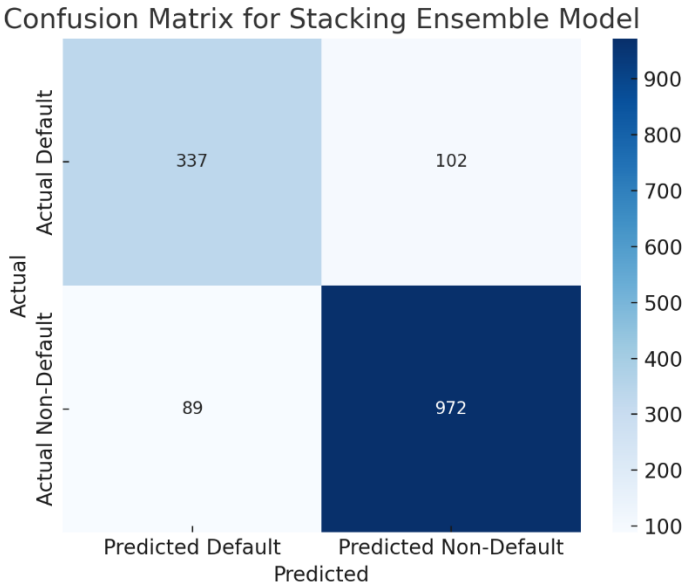

Figure 2: Confusion matrix (stacking ensemble model)

REFERENCES

1. Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1), 111-138.
2. Hamori, S., Kawai, M., Kume, T., Murakami, Y., & Watanabe, C. (2018). Ensemble learning or deep learning? Application to default risk analysis. Journal of Risk and Financial Management, 11(1), 12.
3. He, H., & Fan, Y. (2021). A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction. Expert Systems with Applications, 176, 114899.
4. Kim, H., Cho, H., &Ryu, D. (2020). Corporate default predictions using machine learning: Literature review. Sustainability, 12(16), 6325.
5. Li, W., Ding, S., Chen, Y., & Yang, S. (2018). Heterogeneous ensemble for default prediction of peer-to-peer lending in China. Ieee Access, 6, 54396-54406.
6. Li, W., Ding, S., Wang, H., Chen, Y., & Yang, S. (2020). Heterogeneous ensemble learning with feature engineering for default prediction in peer-to-peer lending in China. World Wide Web, 23(1), 23-45.
7. Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. Mathematics, 8(10), 1756.
8. Li, Y., & Chen, W. (2021). Entropy method of constructing a combined model for improving loan default prediction: A case study in China. Journal of the Operational Research Society, 72(5), 1099-1109.

9. Madaan, M., Kumar, A., Keshri, C., Jain, R., &Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012042). IOP Publishing.
10. Moscatelli, M., Parlapiano, F., Narizzano, S., &Viggiano, G. (2020). Corporate default forecasting with machine learning. Expert Systems with Applications, 161, 113567.
11. Xu, J., Lu, Z., &Xie, Y. (2021). Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1), 18759.
12. Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. Expert systems with applications, 34(2), 1434-1444.
13. Zhou, J., Li, W., Wang, J., Ding, S., & Xia, C. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and its Applications, 534, 122370.
14. Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, 503-513.
15. Zhu, Y., Zhou, L., Xie, C., Wang, G. J., & Nguyen, T. V. (2019). Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. International Journal of Production Economics, 211, 22-33.